GEM STORES CO. LTD

Data Analysis

Siddhi Padekar

29.08.2021

Table of Contents

Executive Summary

Gem Stones co ltd, which is a cubic zirconia manufacturer. Cubic zirconia is an inexpensive diamond alternative with many of the same qualities as a diamond. The dataset contains the prices and other attributes of almost 27,000 cubic zirconia. In this problem statement, we will study how to predict the price for the stone and how to distinguish between higher profitable stones and lower profitable stones to have better profit share

Introduction

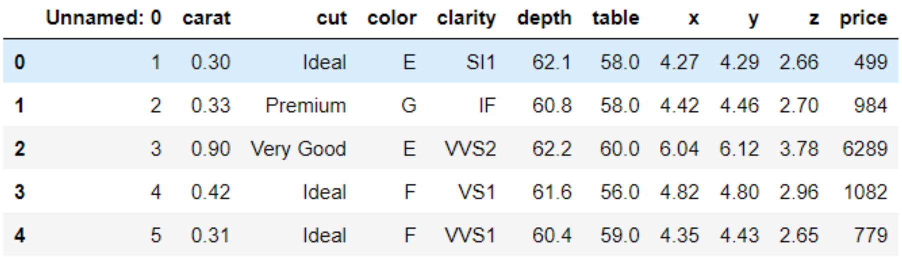
The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset, predict the price for the stone on the bases of the details given in the dataset using linear regression and Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.

Data Dictionary:

Below is the brief description of data set

Variable Name	**Description**
Carat	Carat weight of the cubic zirconia.
Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Colour of the cubic zirconia.With D being the worst and J the best.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = flawless, l1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
Depth	The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	the Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

Sample of the dataset:



	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Above is sample details of data set showing first 5 rows.

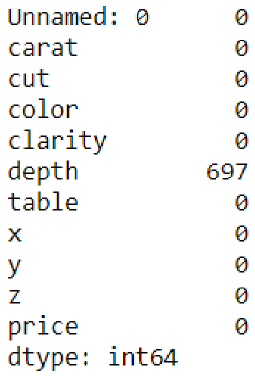
Price is the target variable as we need to predict price of Cubic zirconia and all other are predictor variable

Exploratory Data Analysis

```
Data columns (total 11 columns):
Unnamed: 0     26967 non-null int64
Carat          26967 non-null float64
Cut            26967 non-null object
Color          26967 non-null object
Clarity        26967 non-null object
Depth          26270 non-null float64
Table          26967 non-null float64
X              26967 non-null float64
Y              26967 non-null float64
Z              26967 non-null float64
Price          26967 non-null int64

dtypes:  float64(6), int64(2), object(3)
memory usage: 2.3+ MB
Shape of Data (26967, 11)
```

1. The data set contains 26967 row, 11 columns .
2. In the given data set there are 2 Integer type features, 6 Float type features. 3 Object type features.

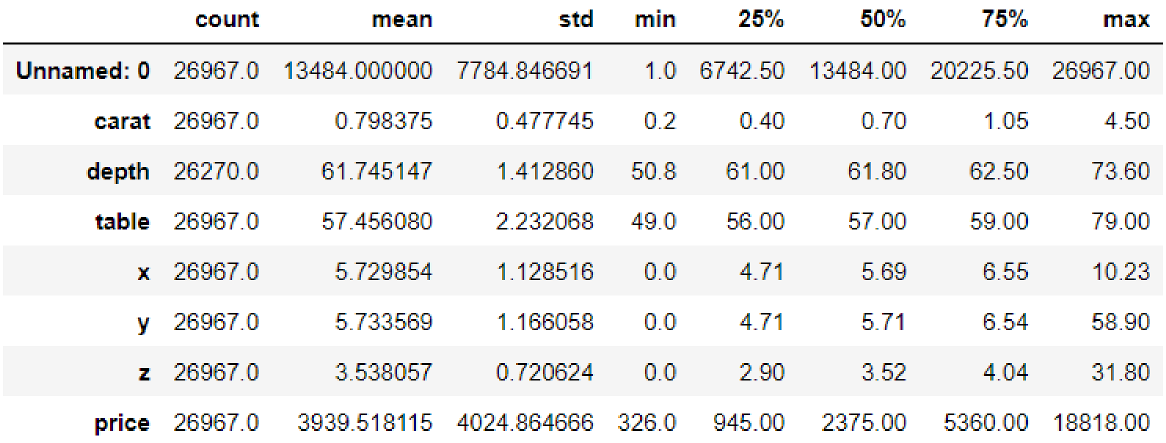
Check for missing values in the dataset:



```
Unnamed: 0      0
carat           0
cut             0
color           0
clarity         0
depth         697
table           0
x               0
y               0
z               0
price           0
dtype: int64
```

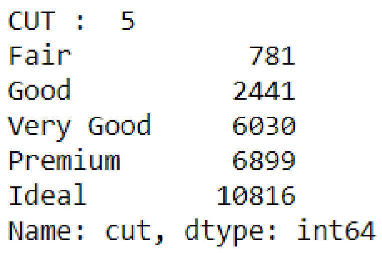
- There are Nul values present only in depth column of data set

Mean, Standard deviation , Min, Max

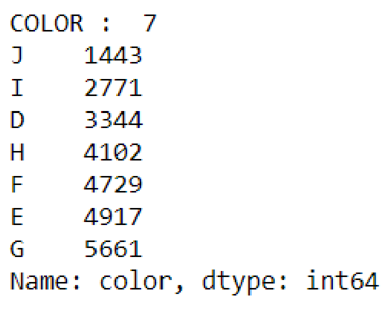


	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.0	13484.000000	7784.846691	1.0	6742.50	13484.00	20225.50	26967.00
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

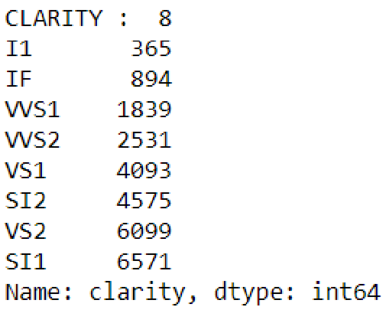
Categorising the object variable (Cut , Color, Clarity)



```
CUT :  5
Fair            781
Good           2441
Very Good      6030
Premium        6899
Ideal         10816
Name: cut, dtype: int64
```

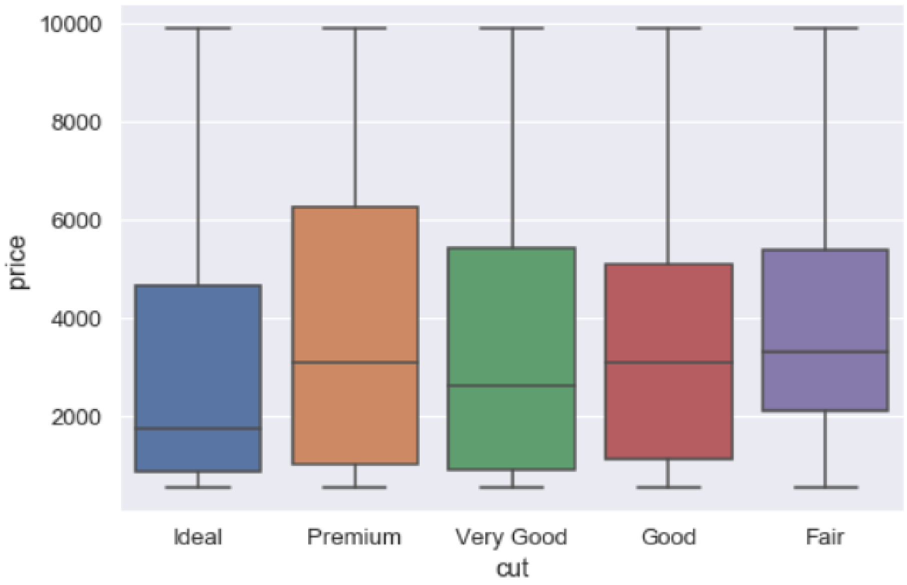


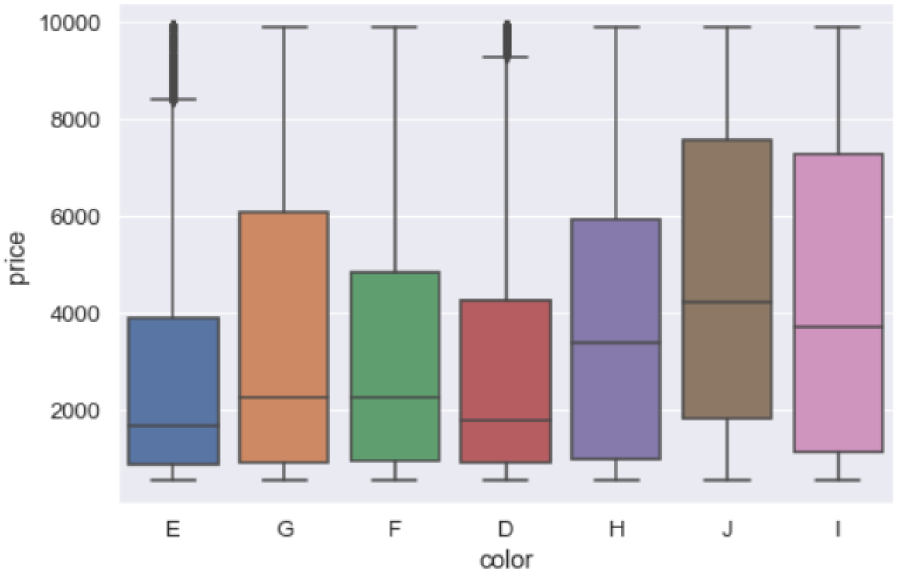
```
COLOR :  7
J     1443
I     2771
D     3344
H     4102
F     4729
E     4917
G     5661
Name: color, dtype: int64
```

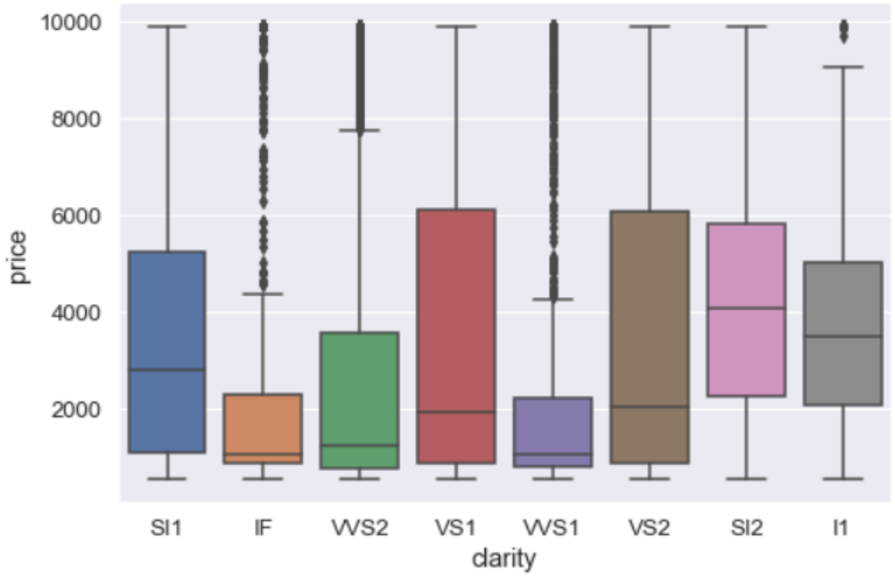


```
CLARITY :  8
I1        365
IF        894
VVS1     1839
VVS2     2531
VS1      4093
SI2      4575
VS2      6099
SI1      6571
Name: clarity, dtype: int64
```

- Cut is categorised into 5 heads – Fair ,Good , Very Good, Premium , Ideal (on the Top is Ideal)
- Color is categorised into 7 heads – J,I,D,H,F,E,G (on the Top is G)
- Clarity is categorised into 8 heads – I1,IF,VVS1,VVS2,VS1, SI2,VS2,SI1 (on the Top is SI1)


10000
8000
6000
4000
2000
price
Ideal
Premium
Very Good
Good
Fair
cut


10000
8000
6000
4000
2000
price
E
G
F
D
H
J
I
color


10000
8000
6000
4000
2000
price
SI1
IF
VVS2
VS1
VVS1
VS2
SI2
I1
clarity

Check for Duplicate values and Removing Duplicate values in the dataset:

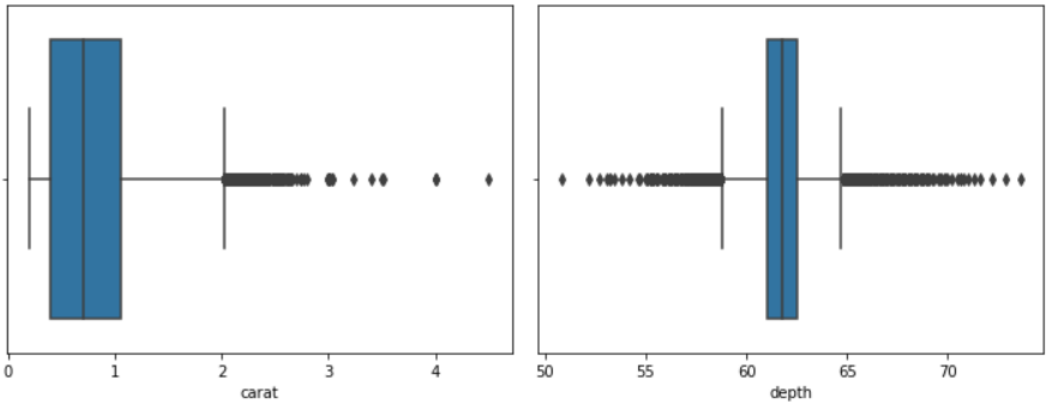
```
Number of duplicate rows = 33

Before removing duplicate values - (26958, 10) – 26958 Rows and 10 Columns
After removing duplicate values- (26925, 10) - 26925 Rows and 10 Columns
```

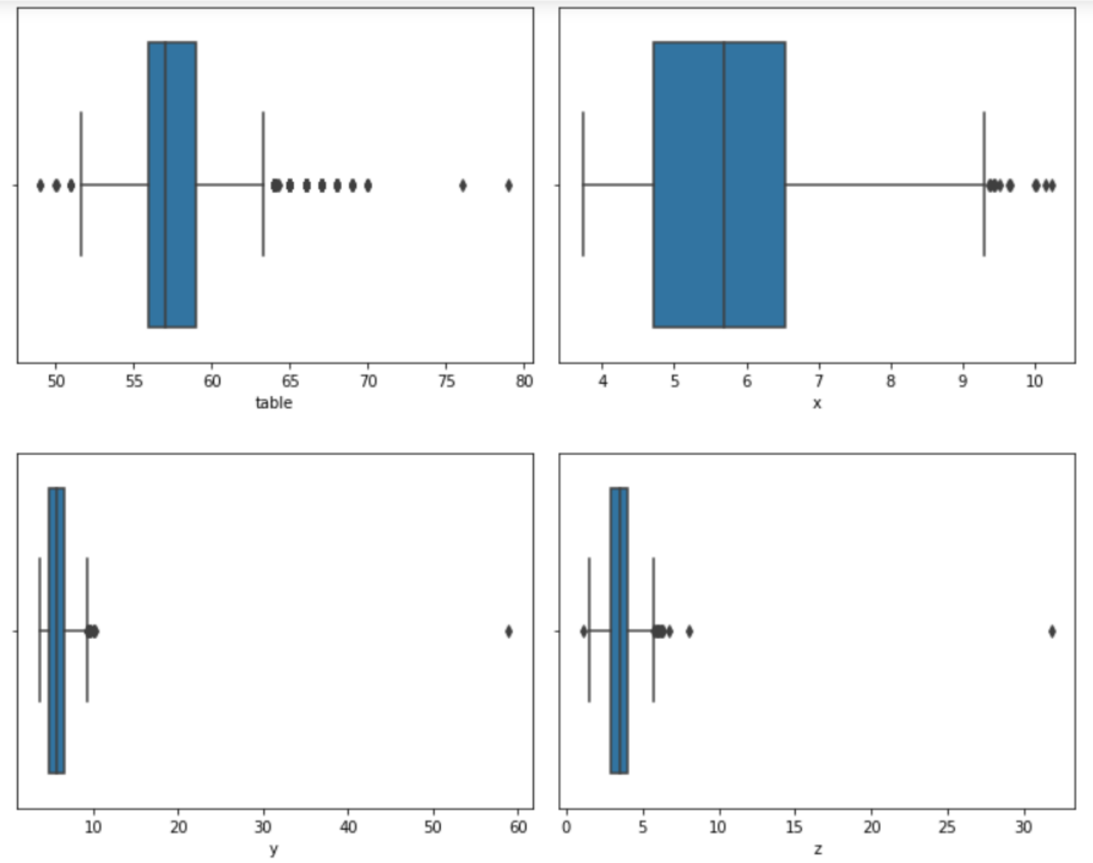
We have dropped duplicate values as they are very small in numbers.

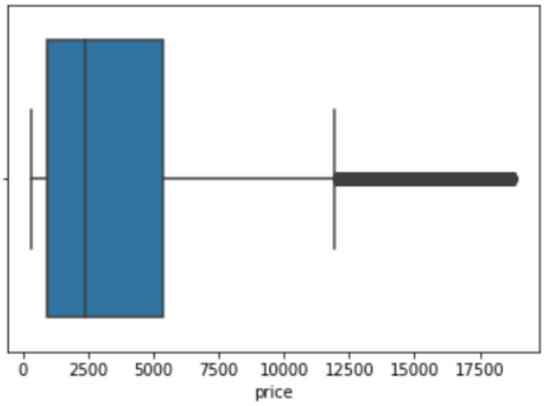
Note- we have dropped first column from the dataset ("Unnamed: 0") as this is only serial numbers

Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.



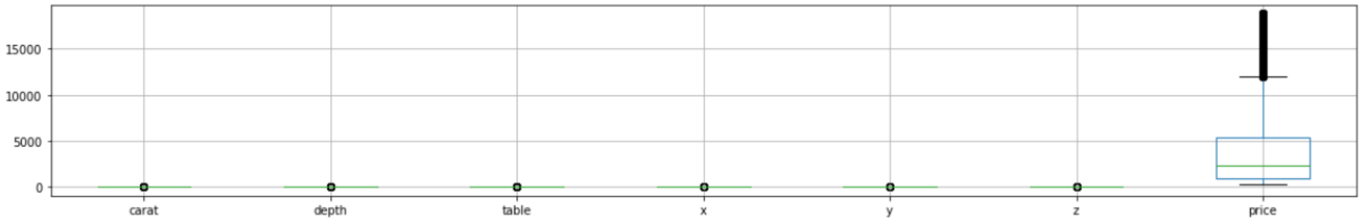




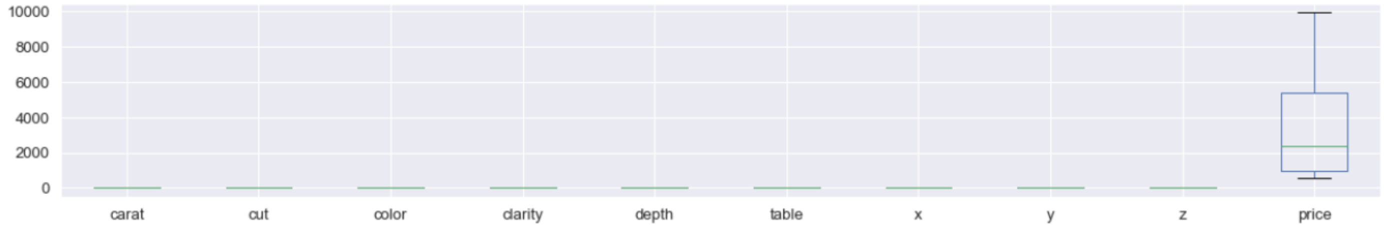
There are significant number of outliers in all the attributes

```
carat --------- 655
depth --------- 1217
table --------- 317
x --------- 12
y --------- 12
z --------- 14
price --------- 1777
```

Before removing Outliers

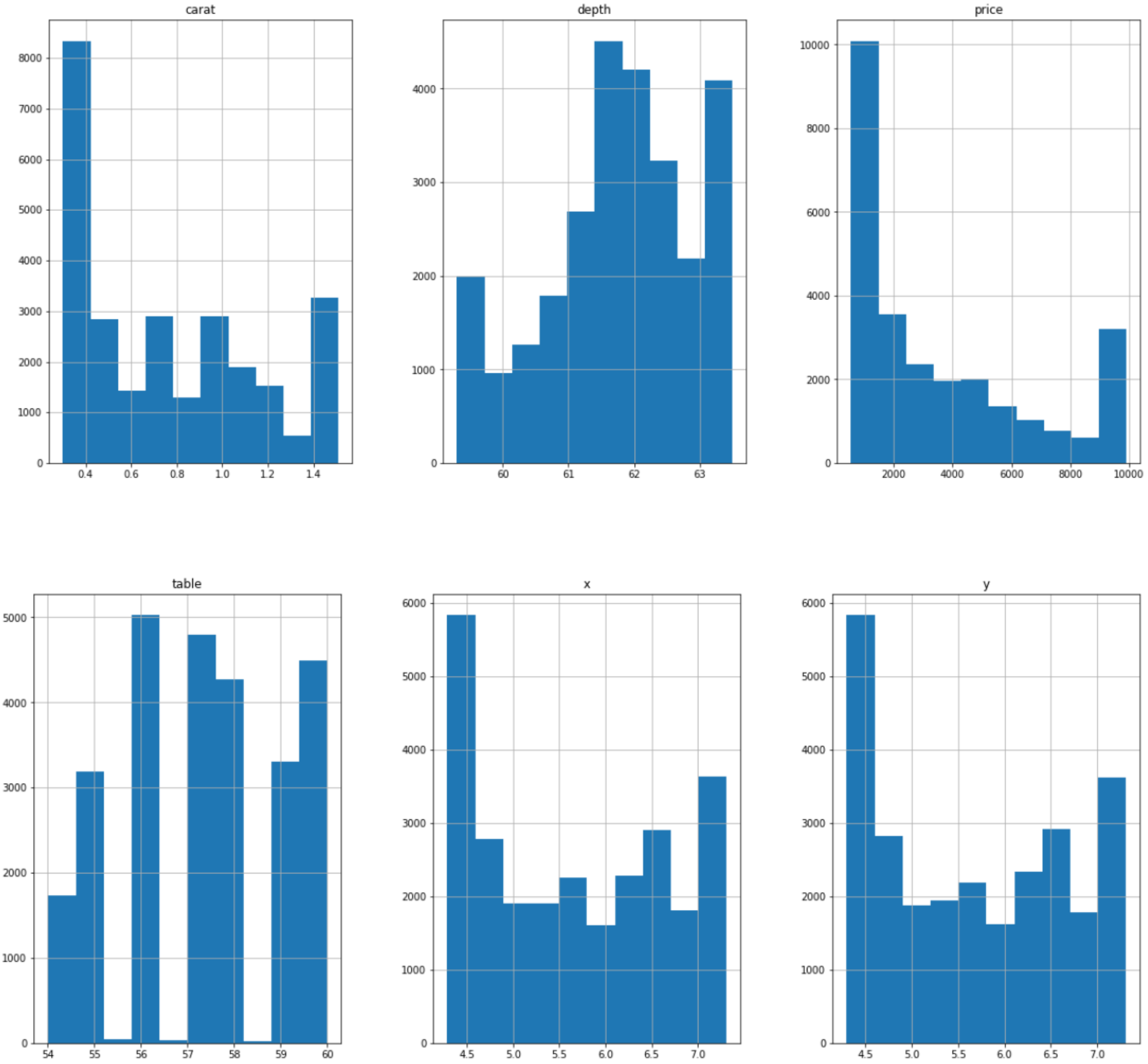


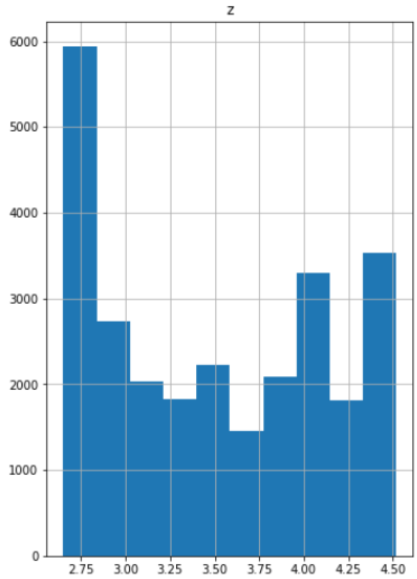
After removing outliers



Univariate Analysis

Histograms

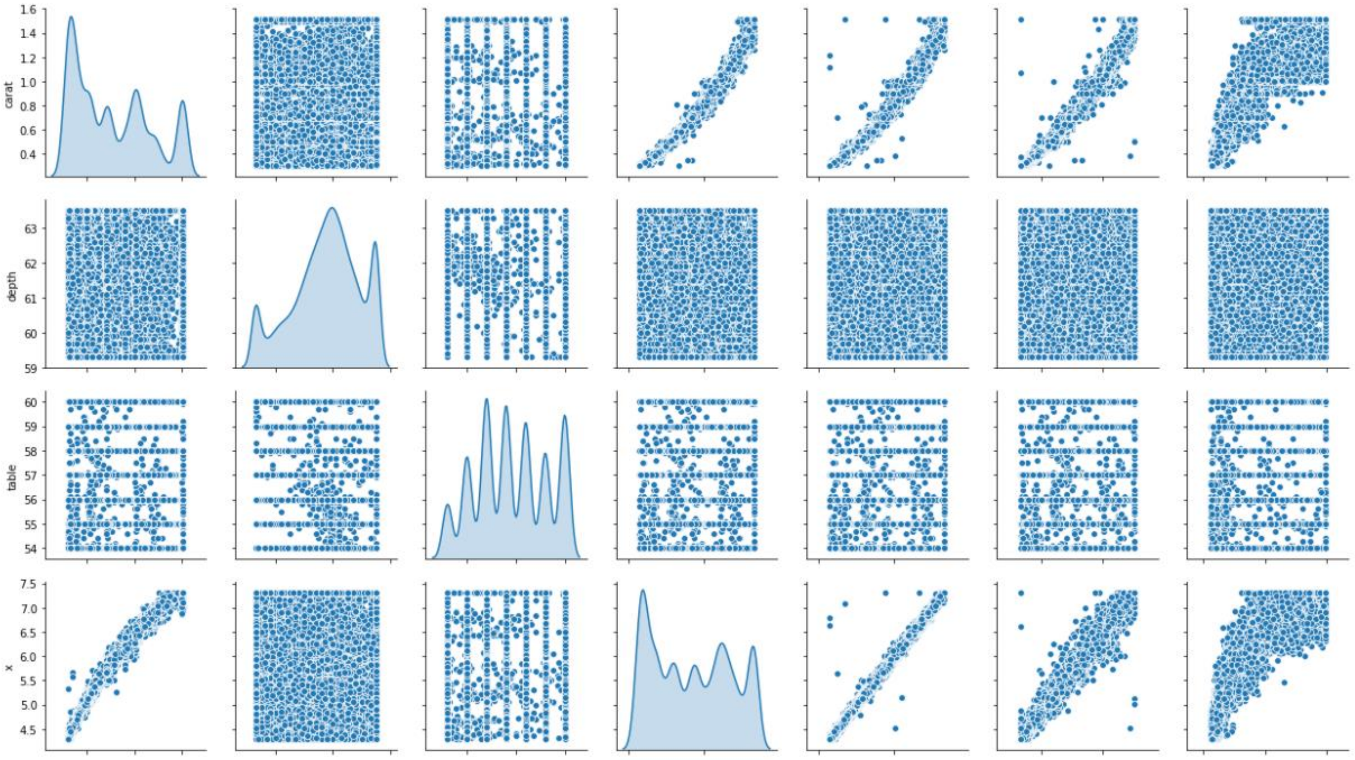


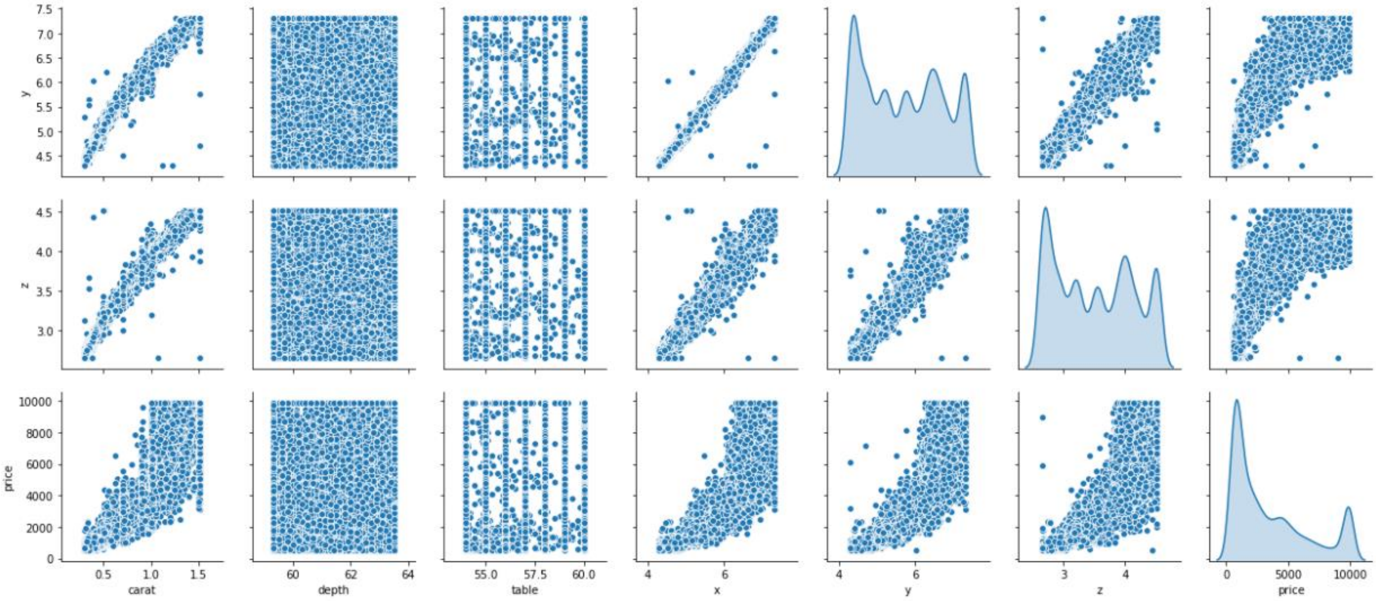


Bivariate Analysis

Pair Plot

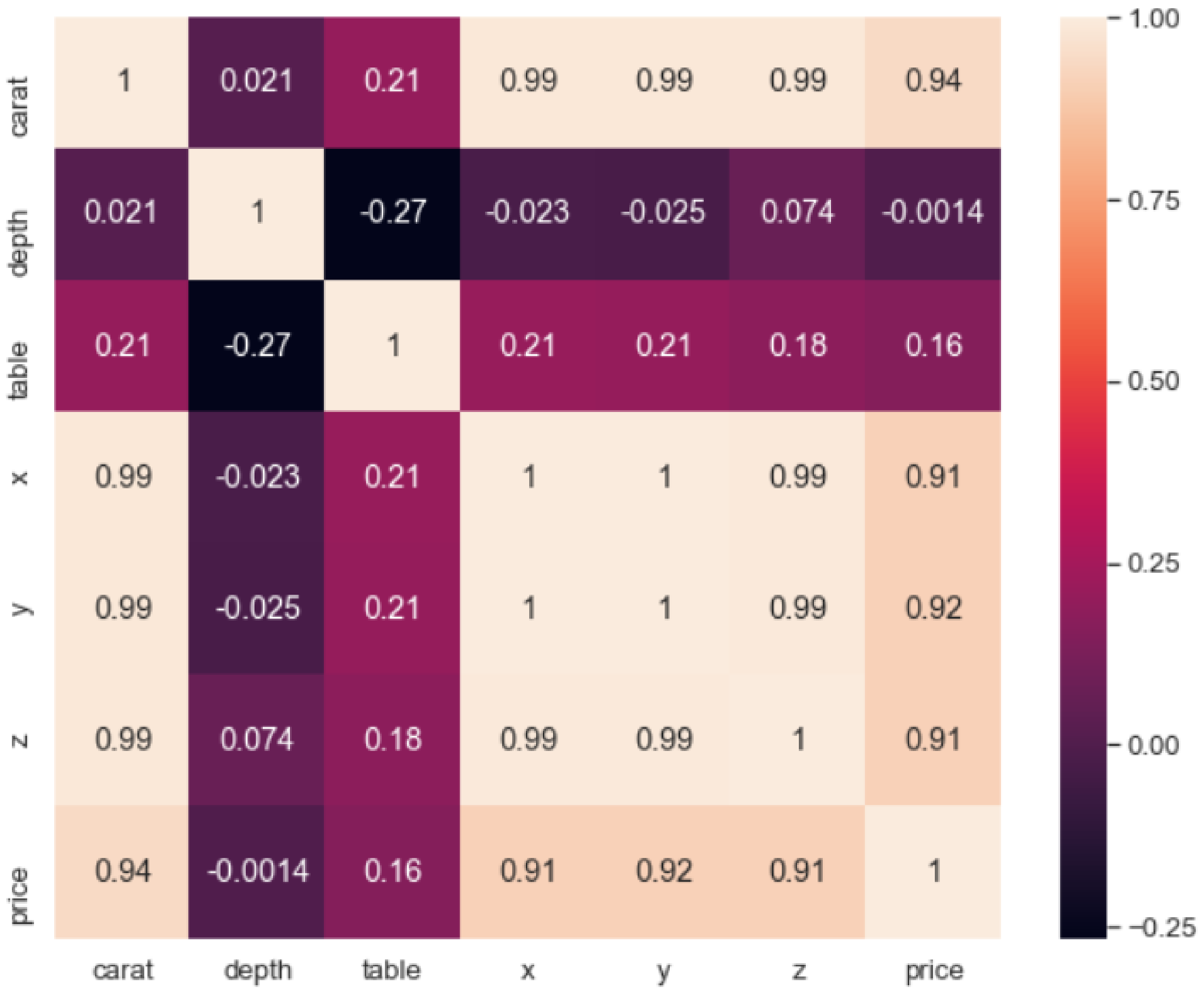
Pair plot visualizes given data to find the relationship between them where the variables can be continuous or categorical.





Multivariate analysis

Correlation Plot



- The variable 'carat', 'x','y', 'z' correlates with target variable 'price' .

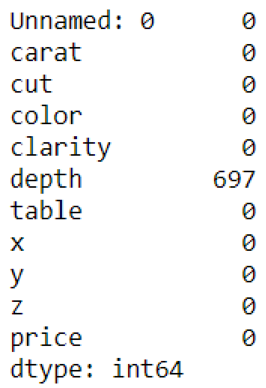
From the correlation plot, we can see that various variable 'carat', 'x','y', 'z' correlates with target variable 'price' . Correlation values near to 1 are highly positively correlated. Correlation values near to 0 are not correlated to each other.

Skewness of every attribute

```
carat     0.502746
depth    -0.408395
table    -0.015381
x         0.132898
y         0.132633
z         0.135138
price     0.917782
dtype: float64
```

We can see that the target feature "price" are heavily "right-skewed".

Imputing Null values



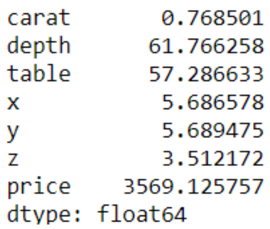
```
Unnamed: 0      0
carat           0
cut             0
color           0
clarity         0
depth         697
table           0
x               0
y               0
z               0
price           0
dtype: int64
```

Exept depth, in all the column there is non null value. Count of null value is 697.

In this case, the mean or median values can be used for imputing the missing or null values of the continuous numerical variables as the mean and median values does not have much difference.

The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the number of values in the set. The median is the middle value when a data set is ordered from least to greatest.

Hence, we have imputed null values with **mean** in the data set.



```
carat       0.768501
depth      61.766258
table      57.286633
x           5.686578
y           5.689475
z           3.512172
price    3569.125757
dtype: float64
```

Are there any values Equal to 0?

Yes there are values equal to 0, however they are faulty data as length , width ,height of any Cubic zirconia gem can not be 0 . Hence we have removed them from the data.

```
Number of rows with x == 0: 3
Number of rows with y == 0: 3
Number of rows with z == 0: 9
Number of rows with depth == 0: 0
```

Do you think scaling is necessary in this case?

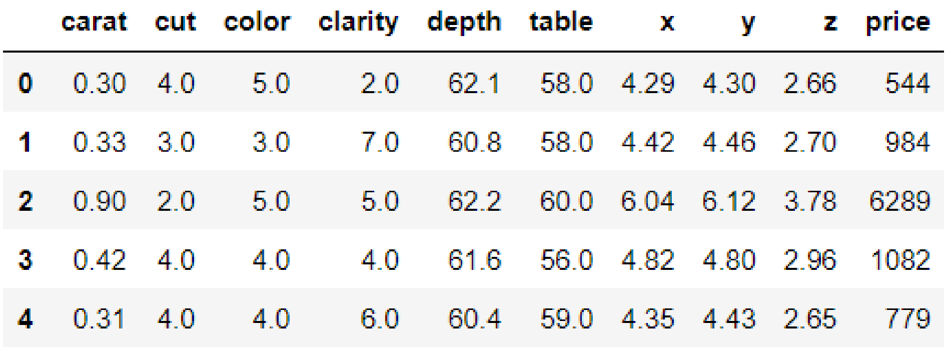
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. Standardization is another scaling technique where the values are centred on the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

In the given data set data set, we can see the all the variable are scaled on different parameters. Like price varies in 1000s unit (Rupees) and depth varies in 100s unit (cms) and table are in 100s unit (percentages) , and carat is in 10s (weights) . Therefore, it is necessary to scale or standardise the data to allow each variable to be compared on a common scale.

However, in this case it is not necessary to scale the data. We will get a solution whether we apply some kind of linear scaling or not. When number of features becomes large, it helps is running model quickly else the starting point would be very far from minima, if the scaling is not done in pre-processing.

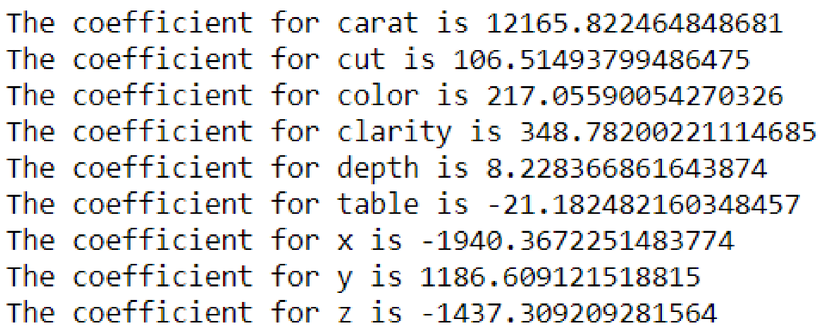
Hence, we have now processed the model without scaling also, we will check the output with scaled data of regression model output.

Sample data set after converting object variable into float.



	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	4.0	5.0	2.0	62.1	58.0	4.29	4.30	2.66	544
1	0.33	3.0	3.0	7.0	60.8	58.0	4.42	4.46	2.70	984
2	0.90	2.0	5.0	5.0	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	4.0	4.0	4.0	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	4.0	4.0	6.0	60.4	59.0	4.35	4.43	2.65	779

Linear regression after splitting data into train and test



```
The coefficient for carat is 12165.822464848681
The coefficient for cut is 106.51493799486475
The coefficient for color is 217.05590054270326
The coefficient for clarity is 348.78200221114685
The coefficient for depth is 8.228366861643874
The coefficient for table is -21.182482160348457
The coefficient for x is -1940.3672251483774
The coefficient for y is 1186.609121518815
The coefficient for z is -1437.309209281564
```

We can obverse that there is a direct relation between following independent variables and price.

A unit increases in carat the price also increases with 12165.82 .

A unit increase in cut the price increases by 106.51

A unit increase in color the price increases by 217.05

A unit increase in clarity the price increases by 348.78

A unit increase in depth the price increases by 8.22

A unit increase in y the price increases by 1186.609

However,

A unit decrease in table the price also decrease with -21.18.

A unit decrease in x the price also decrease with -1940.36

A unit decrease in z the price also decrease with –1437.30

The intercept (often labeled as constant) is the point where the function crosses the y-axis. In some analysis, the regression model only becomes significant when we remove the intercept, and the regression line reduces to Y = bX + error. If X equals 0, the intercept is simply the expected mean value of Y at that value. If X never equals 0, then the intercept has no intrinsic meaning. In scientific research, the purpose of a regression model is to understand the relationship between predictors and the response. If so, and if X never = 0, there is no interest in the intercept. It doesn't tell you anything about the relationship between X and Y.

The intercept for our model is 2142.1746382196725 . This concludes that our other variables are 0 our price is 2142.17 no interest in the intercept. It doesn't tell you anything about the relationship between X and Y. hence we need to make it 0 by doing zscore technique.

Performance Metrics:

R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.The R-squared value R^2 is always between 0 and 1 inclusive.

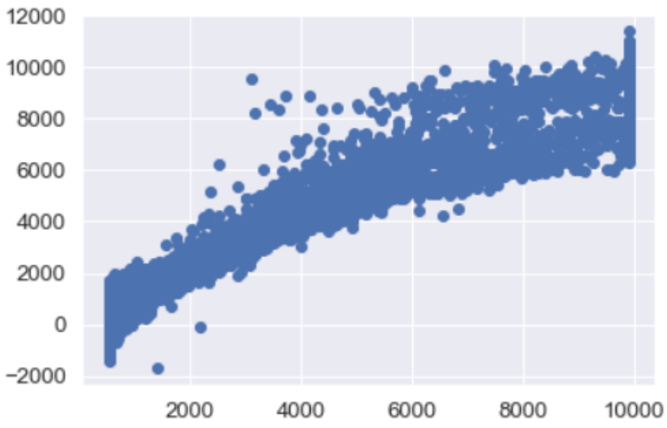
R square on training data is 0.9381845880910002

R square on test data is 0.9390427010465159

0% indicates that the model explains none of the variability of the response data around its mean.100% indicates tha t the model explains all the variability of the response data around its mean. In this regression model we can see the R-square value on Training and Test data respectively 0.938184588091000 and 0.9390427010465159.

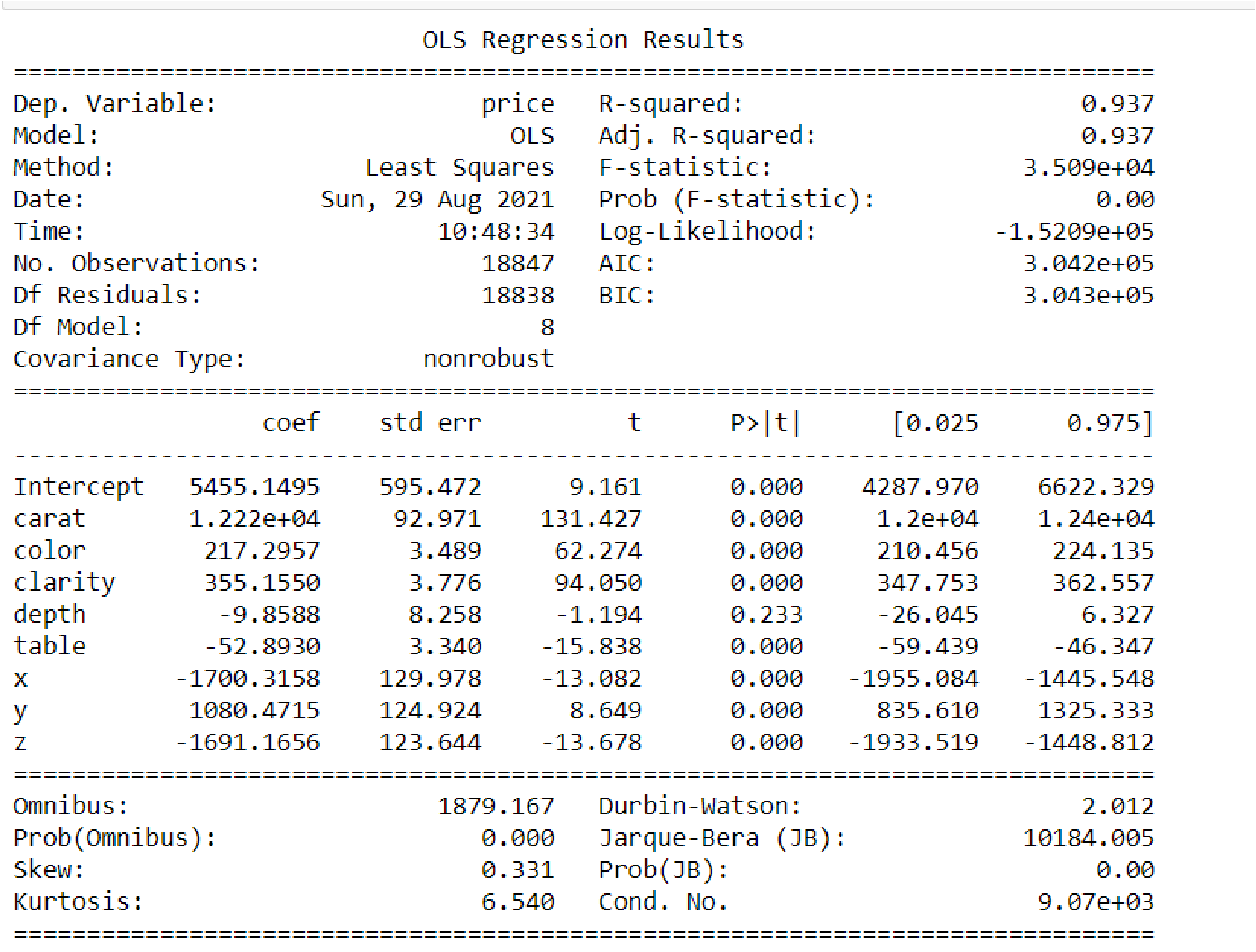
RMSE on Training data is 766.7476323246534

RMSE on Testing data is 767.3860552453051



As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Regression model summary -



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.937
Model:                            OLS   Adj. R-squared:                  0.937
Method:                 Least Squares   F-statistic:                 3.509e+04
Date:                Sun, 29 Aug 2021   Prob (F-statistic):               0.00
Time:                        10:48:34   Log-Likelihood:            -1.5209e+05
No. Observations:               18847   AIC:                         3.042e+05
Df Residuals:                   18838   BIC:                         3.043e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5455.1495    595.472      9.161      0.000    4287.970    6622.329
carat       1.222e+04     92.971    131.427      0.000     1.2e+04    1.24e+04
color        217.2957      3.489     62.274      0.000     210.456     224.135
clarity      355.1550      3.776     94.050      0.000     347.753     362.557
depth         -9.8588      8.258     -1.194      0.233     -26.045       6.327
table        -52.8930      3.340    -15.838      0.000     -59.439     -46.347
x          -1700.3158    129.978    -13.082      0.000   -1955.084   -1445.548
y           1080.4715    124.924      8.649      0.000     835.610    1325.333
z          -1691.1656    123.644    -13.678      0.000   -1933.519   -1448.812
==============================================================================
Omnibus:                     1879.167   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10184.005
Skew:                           0.331   Prob(JB):                         0.00
Kurtosis:                       6.540   Cond. No.                     9.07e+03
==============================================================================
```

If we assume null hypothesis is true, i.e there is no relationship between this variable with price. Now we can ask what is the probability of finding this co-efficient in this drawn sample if in the real world the co-efficient is zero. As we see here the overall P value is less than alpha, so rejecting H0 and accepting Ha that atleast 1 regression co-efficient is not '0'. Here all regression co-efficient are not '0'.

For an example: we can see the p value is showing 0.233 for 'depth' variable, which is much higher than 0.05. That means this dimension is useless. So we can say that the attribute which are having p value greater than 0.05 are poor predictor for price.

Final Regression equation

```
(5455.15) * Intercept + (12218.91) * carat + (217.3) * color + (355.16) * clarity +
 (-9.86) * depth + (-52.89) * table + (-1700.32) * x + (1080.47) * y + (-1691.17) *
 z +
```

1. When carat increases by 1 unit, diamond price increases by 12218.91 units, keeping all other predictors constant.
2. When color increases by 1 unit, diamond price increases by 217.3 units, keeping all other predictors constant.

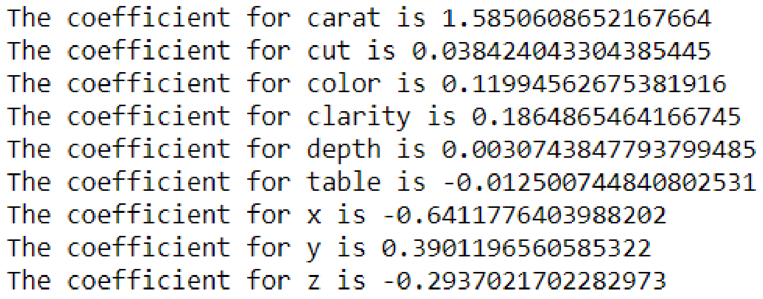
3. When clarity increases by 1 unit, diamond price increases by 355.16 units, keeping all other predictors constant.
4. When y increases by 1 unit, diamond price increases by 1080.47 units, keeping all other predictors constant.

As per model these five attributes that are most important attributes 'Carat', 'Cut', 'color','clarity' and width i.e 'y' for predicting the price.

There are also some negative co-efficient values, for instance, corresponding co-efficient (-1700.32) for 'x', (-9.86) for depth , (- -1691.17) for z and (-52.89) for table This implies, these are inversely proportional with diamond price.

After applying Zscore

Linear Regression



```
The coefficient for carat is 1.5850608652167664
The coefficient for cut is 0.038424043304385445
The coefficient for color is 0.11994562675381916
The coefficient for clarity is 0.1864865464166745
The coefficient for depth is 0.0030743847793799485
The coefficient for table is -0.012500744840802531
The coefficient for x is -0.6411776403988202
The coefficient for y is 0.3901196560585322
The coefficient for z is -0.2937021702282973
```

The intercept for our model is -2.2577061658186334e-16

Inference:

we can see that the from the linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set i.e Unexplained variances on the output.

Impact of scaling:

Now we can observe by applying z score the intercept became -2.2577061658186334e-16. Earlier it was -2142.1746382196725 . the co-efficient has changed and became nearly zero however the overall accuracy still same.

Multi collinearity:

We can observe there are very strong multi collinearity present in the data set.

Recommendations:

The Gem Stones Company should consider the features 'Carat', 'Cut', 'color', 'clarity' and width i.e 'y' as most important for predicting the price.

To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.

As we can see from the model Higher the width('y') of the stone is higher the price.

So the stones having higher width('y') should consider in higher profitable stones.

The 'Premium Cut' on Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.

The Diamonds clarity with 'VS1' &'VS2' are the most Expensive. So these two category also consider in higher profitable stones.

As we see for 'X' i.e Length of the stone, higher the length of the stone is lower the price.

So higher the Length('x') of the stone are lower is the profitability. higher the 'z' i.e Height of the stone is, lower the price.

This is because if a Diamond's Height is too large Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light. That is why

Stones with higher 'z' is also are lower in profitability.